Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-018-00351-6

REGULAR ARTICLE

Exploration of the variability of variable selection based on distances between bootstrap sample results

Christian Hennig[1,2] · **Willi Sauerbrei**[1,3]

Received: 10 November 2017 / Revised: 4 September 2018 / Accepted: 27 December 2018

Abstract

It is well known that variable selection in multiple regression can be unstable and that the model uncertainty can be considerable. The model uncertainty can be quantified and explored by bootstrap resampling, see Sauerbrei et al. (Biom J 57:531–555, 2015). Here approaches are introduced that use the results of bootstrap replications of the variable selection process to obtain more detailed information about the data. Analyses will be based on dissimilarities between the results of the analyses of different bootstrap samples. Dissimilarities are computed between the vector of predictions, and between the sets of selected variables. The dissimilarities are used to map the models by multidimensional scaling, to cluster them, and to construct heatplots. Clusters can point to different interpretations of the data that could arise from different selections of variables supported by different bootstrap samples. A new measure of variable selection instability is also defined. The methodology can be applied to various regression models, estimators, and variable selection methods. It will be illustrated by three real data examples, using linear regression and a Cox proportional hazards model, and model selection by AIC and BIC.

Keywords Linear regression · Cox proportional hazards · Cluster analysis · Multidimensional scaling · Heatmaps

Mathematics Subject Classification 62-07 · 62-09 · 62J20 · 91C15

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11634-018-00351-6) contains supplementary material, which is available to authorized users.

✉ Christian Hennig
christian.hennig@unibo.it; c.hennig@ucl.ac.uk

1 University of Bologna, London, Italy

2 University College London, London, UK

3 Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, London, Germany

Published online: 15 February 2019

1 Introduction

In many regression problems in which the aim is to explain or predict a response y from a set of explanatory variables $x_1, \ldots, x_p$, it is of interest to select a smaller subset of the explanatory variables for fitting a model. Variable selection is done for various reasons:

- A full model with all variables may be ill-conditioned or unstable.
- The practitioner may want a simpler model and a simpler interpretation.
- Prediction can be based on fewer (potentially expensive) measurements.
- There are many uninformative variables in the data set.
- The researcher's main aim may be to find out which variables are relevant influences on y.

It is well known that variable selection can be unstable (Harrell 2001; Sauerbrei et al. 2015); different models (i.e., different choices of explanatory variables) may yield very similar fits of the observations, or quite different fits of the observations by different models may have a similar quality and it may be impossible to tell them reliably apart based on the available data (as is also an issue in other model selection problems, for model-based clustering see Cerioli et al. 2018). Small changes in the data set can result in substantial changes of the selected model. This means that researchers need to be very careful when interpreting the results of variable selection. Particularly, regardless of which technique for variable selection was used, it can be taken for granted neither that the selected variables are all relevant influences on y nor that unselected variables are not relevant influences. For example, if different explanatory variables share similar information about y, it may strongly depend on chance which of them is selected.

The aim of the present paper is to explore the variability of variable selection. It starts from running variable selection methods on bootstrapped data subsets (Sauerbrei and Schumacher 1992; Sauerbrei et al. 2015). The set of models found on different bootstrapped data sets is then explored using distance-based techniques such as multidimensional scaling and cluster analysis. This allows to address issues such as how much variability there actually is, how this variability can be structured and interpreted (i.e., what kind of different models or groups of models deliver very similar fits), how such a structure can be related to the quality of the fits, which observations are fitted differently by different models, which variables make a more or less stable contribution to the models in terms of the resulting fits, to what extent and in what way results from different variable selection methods differ. We also define a new measure of instability in the bootstrap variable selection. Complementary visualisation methods for bootstrap based variables selection are implemented in the mplot package of the statistical software system R, Tarr et al. (2018). Riani and Atkinson (2010) propose a robust variables selection method involving exploration and visualisation of various models, although their aim is not the exploration of the variability of variable selection.

In three examples, we will apply our approach here to variable selection problems in linear regression fitted by least squares, and to Cox-regression in survival analysis. Variable selection is done using backward elimination with Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) as stopping rule. The

ideas can be applied in much more general situations; they extend to different models (generalized linear models, nonlinear regression, classification), different methods of fitting (such as robust or kernel regression), different approaches for variable selection (such as forward or exhaustive search or the Lasso), and the bootstrap can be replaced by subsampling techniques.

In Sect. 2 we briefly introduce the three example data sets. Section 3 introduces the formal methodology, regression and bootstrapped variables selection, dissimilarities, multidimensional scaling (MDS) and clustering. Sections 4, 5 and 6 apply the methodology to the real data sets. This includes the introduction of some helpful scatter- and heatplots in Sect. 4, the comparison of different model selection methods in Sects. 5 and 6, and in Sect. 6 a different model, namely the Cox proportional hazards model for survival data. Section 7 concludes the paper with a discussion.

2 Data sets

We use three data sets to apply and motivate the methodology proposed here. The structure of these data sets differs a lot, which allows us to illustrate different issues. We explore one aspect of model building, namely the decision which variables to include, and assume that the chosen model structure (linear for the first two data sets, Proportional hazards for the third one) is appropriate, which is in line with earlier analyses of these data sets.

The first data set was taken from Rouseeuw and Leroy (1987) and is originally from Coleman et al. (1966). It contains data on $n = 20$ American schools. y is the verbal mean test score, and there are five explanatory variables, namely x_1 (staff salary per pupil), x_2 (percentage of white collar fathers), x_3 (socioeconomic status composition indicator), x_4 (mean teacher's verbal test score), and x_5 (mean mother's educational level). The relevance of selection is debatable given that there are only five variables, but to illustrate various issues an example with a small number of models is ($2^5 = 32$) is suitable. Because of the very small sample size the model selection process is instable and it is likely that models selected in bootstrap samples will differ.

As second example data set we analyse a study on the effects of ozone on school childrens lung growth. Sauerbrei et al. (2015) used this data set as an example for investigating the stability of variable selection using bootstrap. The data set has $n = 496$ observations (children), $p = 24$ variables, and correspondingly $2^{24} = 16{,}777{,}216$ models. For details on the original study see Ihorst et al. (2004), for details on the data set used here see Buchholz et al. (2008). The response is the forced vital capacity (in l) in autumn 1997 (FVC). The explanatory variables are listed in Table 3.

The third data set uses the Cox proportional hazard model for survival times. Krall et al. (1975) analysed the survival times of 65 multiple myeloma patients diagnosed and treated with alkylating agents at West Virginia University Medical Center. There are 16 explanatory variables, which are listed in Table 4.

The response is the rounded survival time in months. Of the 65 patients, 48 were dead at the time of the end of the study, and 17 were alive. With an effective sample size of 48 and $2^{16} = 65{,}536$ models we consider another extreme situation, but this time with the additional issue of censored data. Later we will use observation numbers

and therefore it is useful to know that the observations are ordered in the following way: the first 48 observations are the patients who had died, and within both the died and the surviving patients, observations were ordered from the lowest to the highest survival time (or time in the study after diagnosis).

3 Methodology

Here we give an overview of the formal part of the methodology. Apart from this, a key feature of our analyses are various plots based on MDS and hierarchical clustering. These plots are better introduced in connection with the analysis of the data sets and are therefore presented "on the fly" in the Sects. 4, 5 and 6.

3.1 Regression and bootstrapped variable selection

The general situation of interest here is that the distribution of a response variable $y \in \mathbb{R}$ is a function of variables $x_1, \ldots, x_p$ and a parameter vector β, and the issue of interest is whether this distribution can also be written down as a function of a subset of $\{x_1, \ldots, x_p\}$. With p candidate variables there are 2^p different possible models.

More specifically, before Sect. 6 we assume that we have a data set $\mathbf{Z} = (y_i, x_{1i}, \ldots, x_{pi}),\ i = 1, \ldots, n$, modelled as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i, \tag{1}$$

$e_i \sim \mathcal{N}(0, \sigma^2)$ iid for $i = 1, \ldots, n$ (later we denote $x_{.i} = (x_{1i}, \ldots, x_{pi}), i = 1, \ldots, n$). Usually, the variable selection problem is understood as the task to find $V \subseteq \{1, \ldots, p\}$ so that $j \notin V \Leftrightarrow \beta_j = 0$, although in practice nobody would believe that any true β is exactly zero if it even exists. The present paper is concerned with exploring the variability of variable selection and will therefore neither require that the model holds nor that any β_j is truly zero.

In Sect. 6, we use a Cox proportionals hazard model instead of (1); bootstrapped variable selection is used in the same way as before. The hazard function at time t given the explanatory variables $x_1, \ldots, x_p$ is modelled as

$$\lambda\left(t|x_1, \ldots, x_p\right) = \lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i x_i\right). \tag{2}$$

$\lambda_0(t)$ is nonparametric, and $\beta_1, \ldots, \beta_p$ can be estimated without the need to estimate $\lambda_0(t)$, see Cox (1972).

Given any variable selection method T that returns, for a data set of this kind, a set $\hat{V} \subseteq \{1, \ldots, p\}$ and estimates $\hat{\beta}(V, \mathbf{Z}) = \{\hat{\beta}_j : j \in V\}$, its stability is explored by applying it to b nonparametric bootstrap samples $\mathbf{Z}_i^*,\ i = 1, \ldots, b$ (of same size n with resampling, although other resampling sizes have been used in the literature as well, e.g., Shao 1996) yielding sets $\hat{V}_1, \ldots, \hat{V}_b$ and estimates $\hat{\beta}(V_i, \mathbf{Z}_i^*),\ i = 1 \ldots, b$, see Sauerbrei and Schumacher (1992); Sauerbrei et al. (2015) for a detailed discussion. We

use $b = 500$ for each variable selection method here (for the Coleman data set we only use a single one, for the other two data sets we use AIC and BIC as selection criteria). Apart from the very small Coleman data set, $b = 500$ does not allow to explore all models that could potentially be selected, but unless p is very small, chances are that this is not possible with a substantially larger and computationally realistic b either. To us it seems that $b = 500$ strikes a good compromise between acceptable computing times, a visual structure that can still be explored comfortably by eye, and on the other hand a sufficiently rich coverage of the space of models that makes it quite likely that what is missed will either be further instances of model clusters that are already represented, or quite unlikely "outlying" models.

In the present paper we focus on backward elimination for variable selection, based on Least Squares linear regression for model (1). The stopping criterion for the backward elimination is Akaike's Information Criterion (AIC), i.e., elimination of variables is stopped when the elimination of a variable makes the AIC worse. In Sects. 5 and 6 we will also use the BIC for a demonstration of how the methods introduced here can explore the difference between different variable selection methods in a given data set. See Royston and Sauerbrei (2008) for background.

The dissimilarity-based methodology defined below allows to compare directly the models found in bootstrap samples with the model (or models) $\hat{V}$ found by applying one or more variable selection methods to the full data set. Because these comparisons are of interest, the methodology will be applied to the "model collection" of $B = cb+c$ models, where c is the number of variable selection methods applied to the full data set (in the examples below, either $c = 1$ or $c = 2$). Let $B^* \leq B$ be the number of pairwise different models in the model collection. Note that the same model (in the sense that the same variables were selected) may result from different bootstrapped data sets.

3.2 Dissimilarities between sets of selected variables

Dissimilarity measures between the models found in different bootstrap runs are the main ingredient of our analyses. Many such dissimilarity measures could be constructed. We distinguish two main approaches. A dissimilarity measure can be based on (a) the set of variables in a model or (b) the fitted y-values of the model for all observations in the data set. Both of these are potentially of interest. In some applications the set of variables may be the main focus for interpretation, namely if researchers are mainly interested in finding out what the most important influences on y are. On the other hand, we are also interested in finding out whether the different models result in different groups of fits regarding the predicted values of the observations, and it would be interesting to see to what extent models that are dissimilar in terms of variables are nevertheless similar in terms of the fitted values.

As a dissimilarity measure based on the variables in the model we suggest the Kulczynski-dissimilarity (Kulczynski 1927):

$$d_K(V_1, V_2) = 1 - \left(\frac{|V_1 \cap V_2|}{2|V_1|} + \frac{|V_1 \cap V_2|}{2|V_2|} \right),$$

where $V_1, V_2 \subseteq \{1, \ldots, p\}$ are two subsets of variables and $|V|$ is the number of elements (variables) in a set V of variables. If at least one of $|V_1|$ and $|V_2|$ is 0, it is sensible to set $d_K(V_1, V_2) = 1/2$. There are two main reasons for the choice of d_K. Firstly, it seems appropriate to use a dissimilarity measure that does not rely on joint absences of variables. Often a large number of variables is available and it is obvious that most of them have to be removed for any acceptable model. Also, we expect that normally in real data with many variables only a few variables have a strong effect on the outcome. Several variables may have a rather weak effect and most variables may have hardly any direct effect and may only be associated with the response through correlation with other variables. Two models with one variable each, but different variables, should not be assessed to be very similar based on the fact that nearly all variables are missing from them both. The most popular dissimilarity measure that does not depend on joint absences is the Jaccard distance (Jaccard 1901),

$$d_J(V_1, V_2) = 1 - \left(\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}\right).$$

This has the disadvantage that according to it models with few variables that are nested in much bigger models are far away from these bigger models, which is undesirable because most "fitting work" in the bigger model may be done by the one or few variables that contribute strongest and are therefore most likely to appear also in smaller models. The Kulczynski dissimilarity avoids this issue by relating $|V_1 \cap V_2|$ to both $|V_1|$ and $|V_2|$ rather than $|V_1 \cup V_2|$, which in case of nested models is just the size of the bigger model. This comes at the price that the Kulczynski dissimilarity does not fulfill the triangle inequality (which is why we refer to it as "dissimilarity" rather than as "distance"), as opposed to the Jaccard distance. See Hennig and Hausdorf (2006) for a discussion of this and why it may be seen as an advantage of the Kulczynski dissimilarity in cases like the one considered here.

A reviewer suggested that the Kulcynski-dissimilarity could be modified so that variables are not counted as "joint presences" if they are present in both models but with different estimated regression parameter sign. Whether this is preferable is not an issue of "right" or "wrong" but rather of how the researcher chooses to interpret similarity. In our version, it refers to the idea that the variable is taken as "influential" in a model rather than how exactly the influence plays out. The next section presents another different formalisation of dissimilarity between models.

3.3 Dissimilarities between fits of observations

As dissimilarity measure between the fits from the two models based on the variable sets V_1 and V_2 we suggest the L_1-distance between the vector of fits, i.e.,

$$d_F(V_1, V_2) = \sum_{i=1}^{n} |f_{V_1}(x_{.i}) - f_{V_2}(x_{.i})|, \text{ where}$$
$$f_V(x_{.i}) = \sum_{j \in V} \hat{\beta}_j(V, \mathbf{Z}) x_{ji}, \; i = 1, \ldots, n,$$

$V, V_1, V_2 \subseteq \{1, \ldots, p\}$. Note that in order to make the fits from the different models better comparable, they are refitted on the whole data set (from now on referred to as the "original data set"), i.e., we use $\hat{\beta}(V_i, \mathbf{Z}),\ i = 1, \ldots, B$, rather than $\hat{\beta}(V_i, \mathbf{Z}_i^*)$; this also makes it possible to include the c models obtained from the full data set in the B models in the collection, see above [such a least squares-refit may not be suitable for all variable selection methods, e.g., for regularization techniques combining variable selection and shrinkage such as the Lasso (Tibshirani 1996)]. Models with the same variables resulting from different bootstrap samples are in this way represented by the same regression parameter estimates and corresponding fits, although when computed on the different bootstrap samples that selected the same model, regression parameter estimates and fits would have been different.

The reason for choosing the L_1-distance here is that the overall distance d_F should not be dominated by large individual distances between fits on certain observations if the fits are very similar on most other observations. Such large individual differences should have an impact, but this should not be upweighted compared with smaller distances as it would be by the L_2-distance based on squares.

3.4 Dissimilarities between observations and between variables

For some of the heatplots introduced later in Sect. 4.5, dissimilarities are also required between observations and between variables. These can be defined based on the bootstrap results as well. These dissimilarities are used for setting up hierarchical clusterings that order observations or variables in the heatplot, so the main aim is to allow for a visualisation that makes it easy to spot the main issues, see Sect. 4.5.

Variables can be characterized by sets of bootstrap runs in which they were selected. As dissimilarity measure between variables we propose the Jaccard distance between these sets. The issue that prompted us to suggest the use of Kulczynski above does not apply here; a variable i that appears rarely can be treated as very different from a variable j that appears often, even if the models in which variable i appears are always those in which variable j appears, too. In any case, we treat variables as similar if they tend to appear together in selected models, which is good for the organisation of heatplots of variables against models, but is quite different from measuring their similarity by the absolute value of their Pearson correlation $|\rho|$, see Sect. 4.2 for an example.

Heatplots involving observations in Sect. 4.5 will mainly show residuals, so we will use the Euclidean distance d_E between the vectors of an observation's residuals $r_i[\hat{\beta}(V_k, \mathbf{Z})] = y_i - \sum_{j \in V_k} \hat{\beta}_j(V_k, \mathbf{Z}) x_{ji}, k = 1, \ldots, B^*$ from the B^* selected models (residuals for survival data are defined differently, see Sect. 6).

So overall we use

- the Kulczynski-dissimilarity d_K between $\frac{B^*(B^*-1)}{2}$ pairs of models,
- the L_1-distance d_F between the fit vectors from the $\frac{B^*(B^*-1)}{2}$ pairs of models,
- the Euclidean distance d_E between the vectors of model-wise residuals of the $\frac{n(n-1)}{2}$ pairs of observations,
- the Jaccard distance d_J between the $\frac{p(p-1)}{2}$ pairs of variables (and additionally a correlation-based similarity measure between pairs of variables for comparison in Sect. 4.2).

3.5 Instability measurement

Based on fits and their standard deviation, the bootstrap results allow to define an absolute measure of model selection instability ("absolute" in the sense that its internal calibration makes values comparable between data sets). This can be achieved by comparing the mean variation of residuals between models within observations (which should be low if models are stable) with the mean variation of residuals between observations within models:

$$s^* = \frac{\frac{1}{B}\sum_{j=1}^{B} \mathrm{SD}\left(r_1\left[\hat{\beta}(V_j, \mathbf{Z})\right], \ldots, r_n\left[\hat{\beta}(V_j, \mathbf{Z})\right]\right)}{\frac{1}{n}\sum_{i=1}^{n} \mathrm{SD}\left(r_i\left[\hat{\beta}(V_1, \mathbf{Z})\right], \ldots, r_i\left[\hat{\beta}(V_B, \mathbf{Z})\right]\right)},$$

where SD denotes the standard deviation. This is based on residuals rather than fits because the standard deviation for different observations within the same model should not be governed by the variation between the values of the explanatory variables. Also, we prefer standard deviations to variances (which could have been used to create a measure in the style of the regression R^2) in order to avoid giving the models and observations with the largest within-model or within-observation variations an unduly large influence on the average. Low values imply that residuals between models are rather similar, which means that the variable selection stability is high.

3.6 Multidimensional scaling

MDS is for mapping dissimilarity data in Euclidean space in such a way that the Euclidean distances between observations approximate the original dissimilarities in an optimal manner. We use MDS for visualising the dissimilarity structure (using d_F or d_K as defined above) of the models selected by the bootstrap replicates.

There are various MDS techniques, see, e.g., Borg et al. (2012). We use ratio MDS here, computed by the R-package smacof (de Leeuw and Mair 2009), which is defined, for a target dimensionality q, by choosing a matrix of Euclidean points $\mathbf{Z} = (z_1', \ldots, z_n')'$, $z_i \in \mathbb{R}^q$, $i = 1, \ldots, n$ in such a way that the Euclidean distances $d_{ij}(\mathbf{Z}) = \|z_i - z_j\|_2$ and a constant $b > 0$ minimize the normalized stress

$$S = \sqrt{\frac{\sum_{i<j}\left(b\delta_{ij} - d_{ij}(\mathbf{Z})\right)^2}{n(n-1)/2}}$$

under the side condition that $\sum_{i<j} b\delta_{ij}^2 = \frac{n(n-1)}{2}$, where $\delta_{ij} = d_F(V_i, V_j)$ or $d_K(V_i, V_j)$, $i, j = 1, \ldots, n$, are the dissimilarities to be approximated.

This means that the Euclidean distances on $\mathbf{Z}$ approximate a normalized version of the original dissimilarities in the sense of least squares. We chose this version because we constructed the dissimilarities in such a way that their values and their ratios should reflect how dissimilar the models are in a well-defined numerical sense (this does not necessarily require that the dissimilarities fulfill the triangle inequality).

Other popular versions of MDS only represent the order of the dissimilarities, or some nonlinear transformation of it, or a linear transformation that doesn't necessarily map zero on zero, all of which are less appropriate here. Classical MDS can be thought of as approximating squared dissimilarities, which gives large dissimilarities too much of an influence on the MDS configuration. For details see Borg et al. (2012).

Obviously, for straightforward visualisation, $q = 2$ is most useful. One should however be concerned about the information loss when representing dissimilarities in low dimensions. The normalized stress, which has a straightforward interpretation in terms of the percentage approximation error, can be used for this. The smacof package also produces a Shepard plot that allows to assess the fit by looking at δ_{ij} versus $d_{ij}(\mathbf{Z})$ (not shown in the examples). We will in the following only show MDS-plots with $q = 2$ because inspection of more dimensions for the data examples treated here did not show further interesting structure. But in general we advise to consider S, Shepard plots and further dimensions of a higher dimensional MDS-solution.

3.7 Clustering

For the exploratory analysis of the models selected by the bootstrap replicates, dissimilarity-based clustering can make the following contributions: (a) it can complement the MDS by using information in the dissimilarities that may be lost by the MDS rendering of the data in low-dimensional Euclidean space; (b) as opposed to visual clustering, it produces well defined and formally reproducible clusters (which of course may coincide with the visually found ones, in which case it can confirm the subjective impression from the MDS); (c) clustering outcomes serve well for ordering the rows and columns of heatplots, see Sect. 4.5.

There is a very large array of clustering methods. We prefer hierarchical methods here because it may be useful to look at groupings at various levels of coarseness (we are not concerned with estimating any "true number of clusters"), and because such methods give more information for structuring heatplots than methods that only produce a partition. We have good experiences in this situation with average linkage (UPGMA) hierarchical clustering, which often is a good compromise between single linkage (which respects gaps but may join cluster with too large within-cluster dissimilarities too early) and complete linkage (which keeps within-cluster dissimilarities low at the expense of at times ignoring flexible cluster shapes). See Hennig (2015) for some considerations regarding the choice of a cluster analysis method.

A general issue is whether analyses should be based on the $B = cb+c$ bootstrap/full model runs or the B^* found models. For data sets with many variables and rather unstable variable selection as the Ozone and Myeloma data sets in Sects. 5 and 6 this does not make much of a difference because B^* is often not much smaller than B. For the Coleman data set (Sect. 4), though, $B^* = 17$ (of 32 possible) and $B = 501$. The computation of the Kulczynski and fit-based dissimilarity measures between models are not affected by this decision; identical models will just produce identical rows in a $B \times B$-dissimilarity matrix. But MDS and the outcome of most clustering methods (but not single and complete linkage hierarchical clustering) will differ depending on whether they are based on dissimilarities between B^* pairwise different objects or

Table 1 Number of selections of variables of Coleman data out of 501 models in the collection

x_1	x_2	x_3	x_4	x_5
301	222	499	440	244

between B objects many of which are identical, in which case they are more strongly influenced by models that were found more often. This may or may not be seen as appropriate; we decided to base MDS and the Jaccard distance between variables on all the B bootstrap replicates (so that all information is used) but to use only the B^* pairwise different models for the heatplots (see Sect. 4.5) and the model clusterings used in them (because this makes it easier to appreciate the models that were not often found in the plots), as well as for the Euclidean distances between observations' residuals.

4 Coleman data

4.1 Regression and bootstrapped variable selection

For the Coleman data, with such a low number of variables, in principle one could include all possible subsets of the variables in an analysis, but we stick to the set of models selected by the $b = 500$ bootstrap replications in order to be consistent with what we recommend in a general case; models not selected by any bootstrap run are quite bad here and not very relevant.

Applying backward elimination with the AIC stopping rule to the original data set selects x_1, x_3 and x_4 with $R^2 = 0.901$. This is the model with the best AIC among all possible models. The collection of $B = 501$ models yielded $B^* = 17$ different models in this data set, 104 of which yielded the model that is best on the original data (one model was found more often, the full model with all five variables was kept 139 times). It is also of interest how often the variables appeared in the selected models, which is indicated in Table 1. All the variables appeared in more than 40% of the models, and one may wonder whether these variables were pre-selected out of a bigger, not publicly available set.

The bootstrap variable selection instability is $s^* = 0.235$, the residual variation between models within observations is about a quarter of the variation between observations within models.

4.2 Dissimilarities

With $B^* = 17$ we have $(17\times16)/2$ dissimilarities between models. It is surprising to see that in some of the replications the full model with 5 variables and in other replications two different univariate models were selected, see Fig. 2, which also shows that one of the univariate models is clearly outlying.

The left side of Fig. 1 shows a scatterplot of the two different distances, d_K and d_F. It shows that for the Coleman data the similarity of some pairs of models is assessed quite differently by d_K and d_F; among the smallest Kulczynski dissimilarities d_K (i.e.,

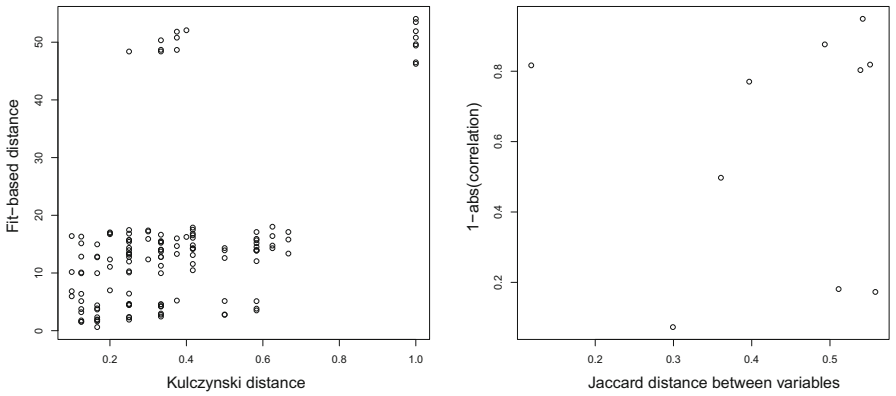

Fig. 1 Left side: Scatterplot of Kulczynski versus fit-based dissimilarities for the Coleman data set. Right side: Scatterplot of Jaccard distance versus $1 - |\rho|$ for pairs of variables in the Coleman-data set

Table 2 Pearson correlation matrix for Coleman data

	x_1	x_2	x_3	x_4	x_5
x_1	1.00	0.18	0.23	0.50	0.20
x_2	0.18	1.00	0.83	0.05	0.93
x_3	0.23	0.83	1.00	0.18	0.82
x_4	0.50	0.05	0.18	1.00	0.12
x_5	0.20	0.93	0.82	0.12	1.00

pairs of models with very similar set of variables) are models with d_F (distance of fits) up to about 18, much larger than the smallest d_F-values, indicating that inclusion or exclusion of a single variable can cause quite a difference in fits. There is also a group of very high d_F around 50 corresponding to moderate d_K around 0.4, which looks like a distinct cluster, and another one with d_F high and $d_K = 1$. Both of these "clusters" refer to dissimilarities involving the model with only the "white collar fathers" variable in it (see Fig. 4 discussed later); some of the other models have no variable in common with this ($d_K = 1$) and some have one or more variables in common, but still the fits are very different. On the other hand, some of the pairs of models with smallest d_F have d_K up to 0.6. Figure 4 as introduced below shows some information about which variables make a bigger difference in terms of fits.

The right side of Fig. 1 shows that dissimilarity assessment between pairs of variables by d_J is quite different from measuring their similarity by the absolute value of their Pearson correlation $|\rho|$ as mentioned in Sect. 3.2; the correlations are given in Table 2. This means that the property of being selected together for two variables in this data set has little relation to their correlation, which particularly means that strong correlations are not a main driving force for variable selection here.

4.3 Multidimensional scaling

The MDS-solution for the Coleman data with fit-based distance is shown on the left side of the Figs. 2 and 3, together with some further visualisation elements. In Fig. 2,

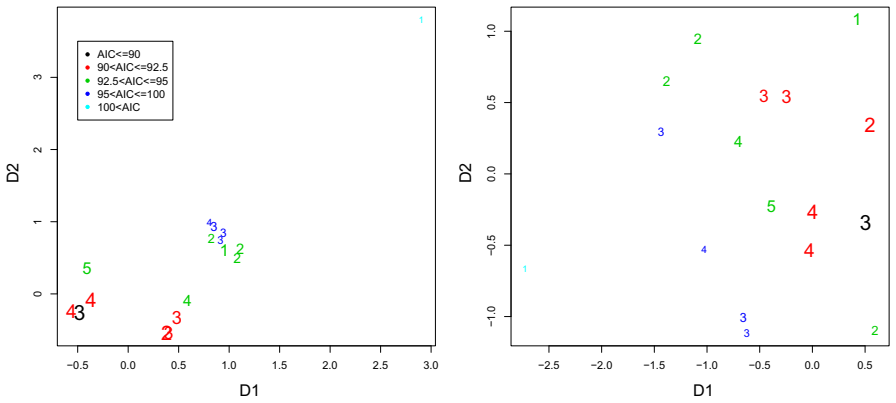

Fig. 2 Left side: 2-dimensional MDS solution for the fit-based dissimilarity between the 17 models for the Coleman data set. Colors indicate how good a model was according to the AIC (black indicates the optimal model according to the AIC); symbol sizes are proportional to the rank (i.e., the biggest symbol indicates the best model according to the AIC). The numbers indicate how many variables there are in a model. Right side: Same, but with MDS solution for the Kulczynski dissimilarity (color figure online)

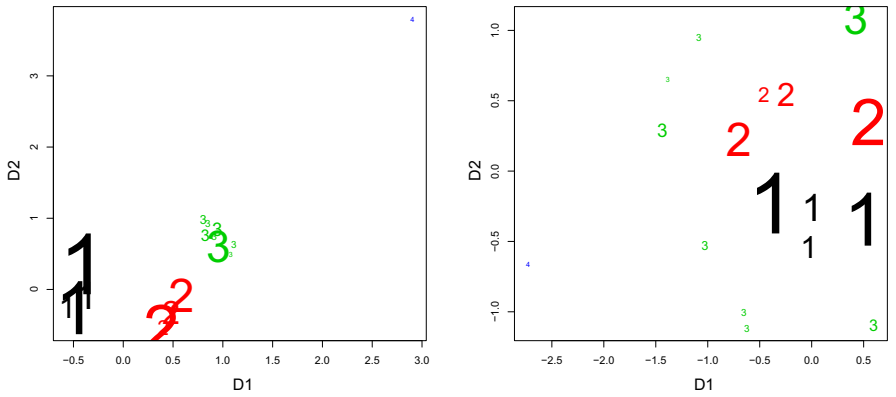

Fig. 3 Left side: 2-dimensional MDS solutions for the fit-based dissimilarity between models for the Coleman data set with 4-cluster average linkage clustering from fit-based similarity. Sizes of symbols are proportional to the square root of how often a model was found. Right side: same with MDS solutions for the Kulczynski dissimilarity between models (color figure online)

the numbers indicate how many variables are in the models, and the colors and number sizes indicate how good the model is according to the AIC.

On the left side, there seem to be four "clusters" of models in this plot, one of which is just the single outlying model with apparently vastly different fits, with the worst AIC-value (light blue). This model was only found once (in Fig. 3, the sizes of the numbers indicate how often a model was found; the numbers and colors there refer to the clusters, see Sect. 4.4) and has only a single variable, namely percentage of white collar fathers, as can be seen in Fig. 4. Close to the middle of the plot there seems to be a group of models with similar fits that are far from optimal according to the AIC. The two groups of fits on the lower right side were selected most often (cluster 1 selected 290 times, cluster 2 selected 151 times, cluster 3 selected 59 times, cluster 4 selected

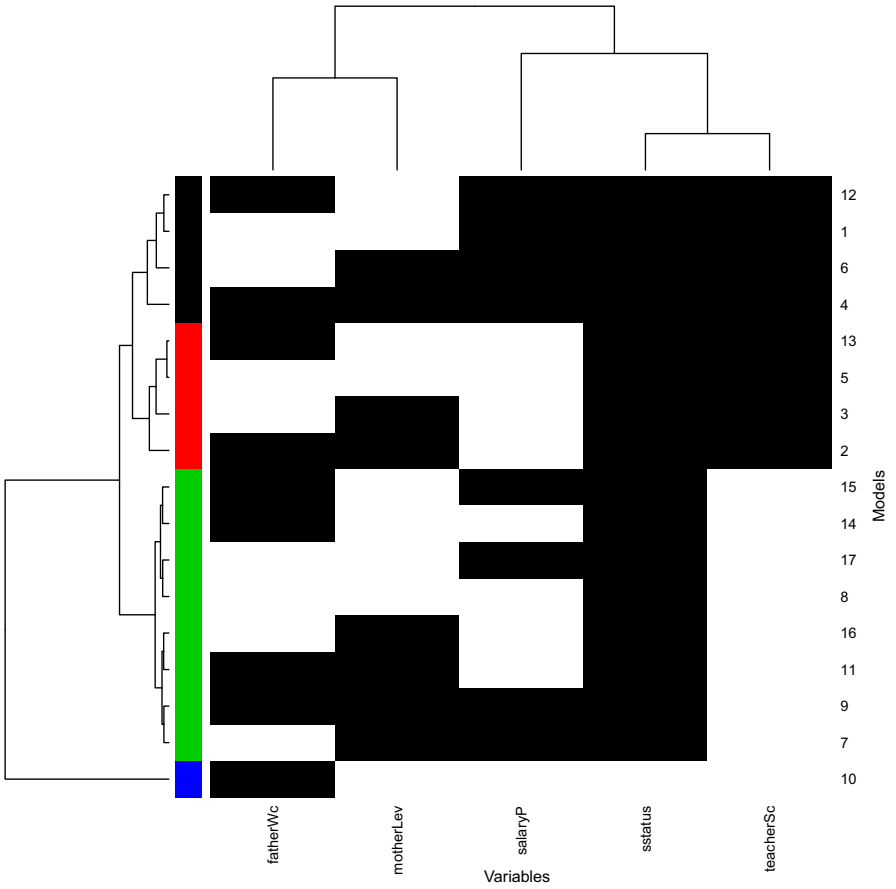

Fig. 4 Variables in models for Coleman data; model clustering dendrogram from fit-based distance. Colors on the left side correspond to clusters from Fig. 3 (black—1, red—2, green—3, blue—4) (color figure online)

once out of $B = 501$) and yield the best fits. In Sect. 4.5 we will explore these groups of fits in more detail in order to find out how these different models interpret the data differently.

The MDS-solution with $q = 2$ yields $S = 0.118$ for the fit-based distance, which according to experience is fairly good but not excellent.

The right sides of the Figs. 2 and 3 show the 2-dimensional MDS-solution for the variable-based Kulczynski dissimilarity, with the same meaning of the additional plot elements as on the left side.

This shows the same "outlier model" as before, now far on the left side. As the space of subsets of $\{1, 2, 3, 4, 5\}$ is quite discrete, the other models are not so clearly clustered. One thing that can be seen on the right side of Fig. 2 is that the models with the next best AIC values (red) are close to the best model (black) also in terms of the Kulczynski dissimilarity. The model with all five variables is central regarding the Kulczynski dissimilarity; it belongs to the black cluster 1. Models with fewer variables are more marginal.

4.4 Clustering

For the Coleman data with fit-based distance between models, the average linkage dendrogram cut at four clusters corresponds with an intuitive visual clustering, see the left side of Fig. 3. The right side of Fig. 3 shows the same clustering on the MDS plot based on the variable-based Kulczynski dissimilarity.

In the Kulczynski-MDS plot (Fig. 3, right side), clusters 1 and 2 are more central and not separated; and cluster 3 looks rather heterogeneous regarding the variables. Connecting this with Fig. 2, the models with 1 and 2 variables in this cluster are actually better, regarding the AIC, than those with 3 and 4 variables, although models in clusters 1 and 2, which are superior in terms of the AIC, tend to have more variables. The AIC is known to favor rather "big" models, see also Sects. 5 and 6.

The next important issue is to explain the clustering, i.e., what characterizes the different fits, and which variables are important for determining to which cluster a model belongs. This can be done using heatplots.

4.5 Heatplots

Heatplots are probably the most useful tool for visualising the bootstrap results. We use them here for showing the variables in all the bootstrapped models (Fig. 4) and for analysing the fits from the different models (Figs. 5, 6). In Figs. 5 and 6, grey scales correspond to the raw fits by the models as indicated by the Color Key. Differences between models are usually more pronounced when looking at residuals (Fig. 6).

Heatplots are ordered by the average linkage clusterings from the fit-based distance (models), the Jaccard distance (variables) and the Euclidean distance on vectors of residuals (observations). This easily allows to connect the heatplots with the cluster structure of the models that was highlighted above. Colors on the left side of the heatplot correspond to those used in Fig. 3. Cluster 1 in Fig. 3 comprises the models no. 12, 1, 6 and 4, cluster 2 comprises models no. 13, 5, 3 and 2, cluster 4 only has model 10 and the remaining models belong to cluster 3. Figure 4 shows that cluster 1 and 2 on one hand and clusters 3 and 4 on the other hand are distinguished by whether or not x_4 (teacher's test score) is in the model or not. Figures 5 and 6 show considerable differences between the fits of these two groups of clusters. Clusters 1 and 2 are distinguished by the presence or absence of variable x_1 (salary per pupil). Figure 5 shows some specific differences between their fits, which are not as pronounced as between them and cluster 3 and 4.

Model 10 (cluster 4) uses only variable x_2, which makes it outlying, and its fits are very different from all the other models. There is a minority of observations on which its fit is actually the best (e.g., no. 7 and 18), but on the majority it does badly (e.g., no. 11 and 15).

If required, the differences in fits can be interpreted in terms of the specific observations. E.g., observation no. 6 is fitted by a much larger value in cluster 3 (green) than in the other three clusters. Its residual in clusters 1 and 2 (black and red) is around zero, somewhat worse in cluster 4 and quite high in cluster 3.

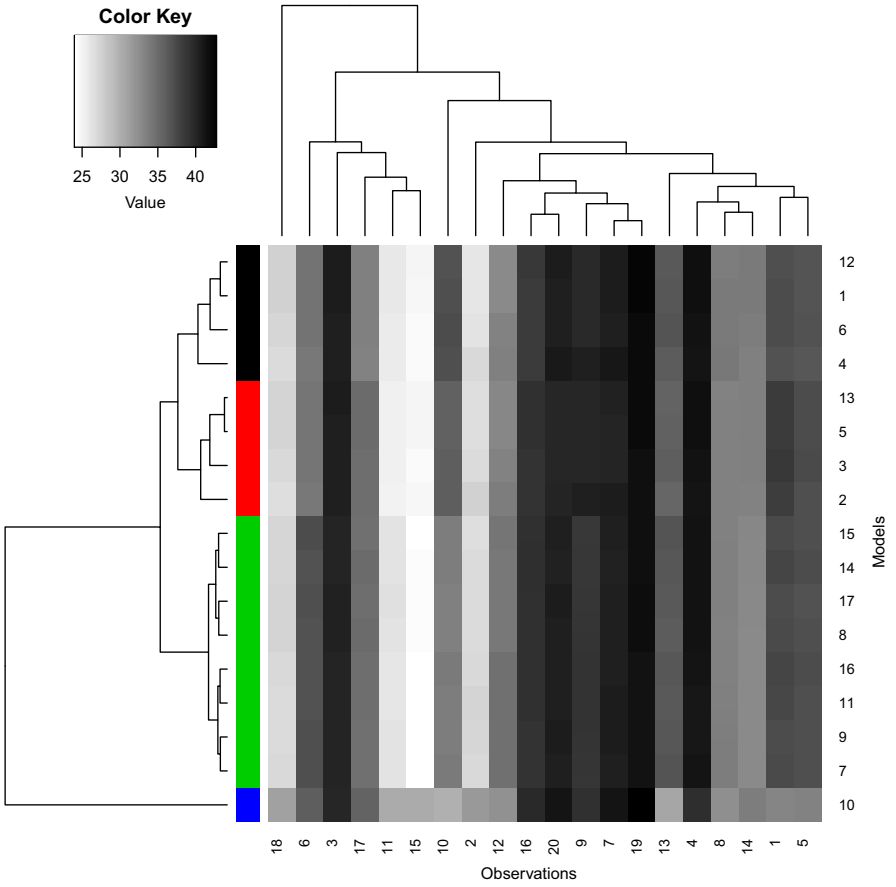

Fig. 5 Model fits of observations for Coleman data; model clustering dendrogram from fit-based distance. Colors on the left side correspond to clusters from Fig. 3 (black—1, red—2, green—3, blue—4) (color figure online)

The models with the best AIC-values are in clusters 1 and 2 (Fig. 2), but looking at Fig. 6 it can be seen that the models in cluster 3 by and large produce lower absolute values of residuals (i.e., color closer to white) for more observations than the models in clusters 1 and 2, which makes these models attractive from a robust perspective. The models in clusters 1 and 2 deliver much better fits for observations 6 and 10, which account largely for the better AIC-values of these models. This means that these models are not as clearly better as the AIC suggests. The fits in cluster 3 look just as legitimate from this perspective.

Overall the four clusters of models clearly refer to quite different ways to fit the observations, and the heatplots in Figs. 5 and 6 allow to explore the differences on the level of individual observations, here showing that the same data may be seen as supporting also a way of fitting the data that is quite different from the AIC-optimal model 1.

In order to save space, we will omit some analyses for the further examples.

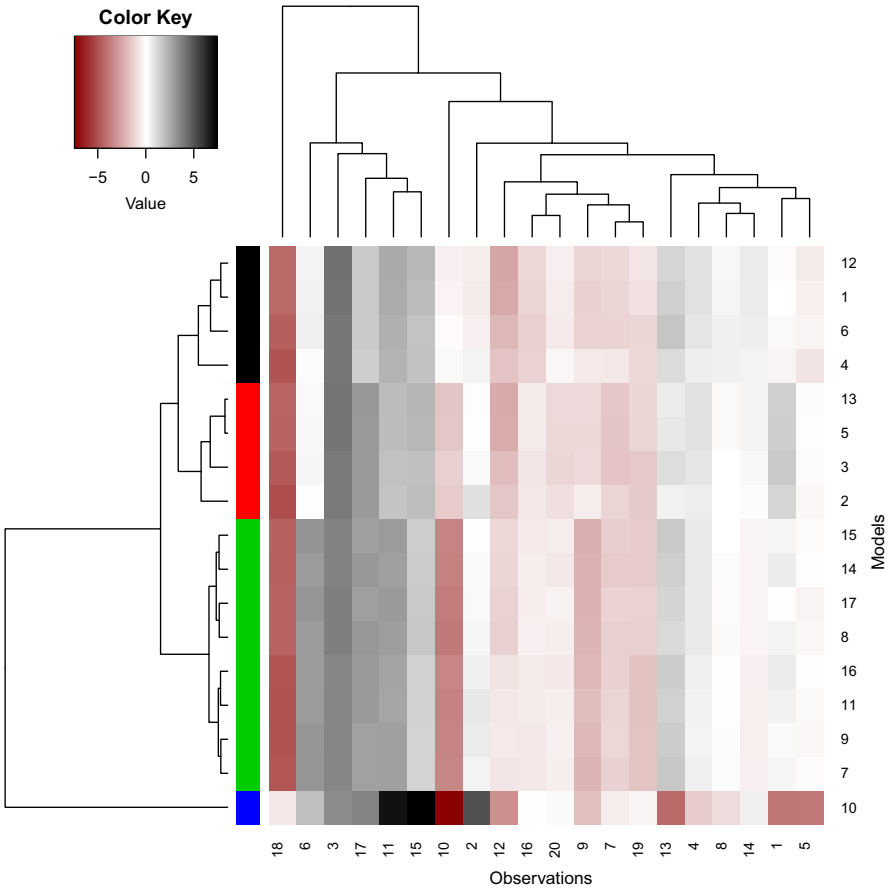

Fig. 6 Residuals for Coleman data with negative residuals in red; model clustering dendrogram from fit-based distance. Colors on the left side correspond to clusters from Fig. 3 (color figure online)

5 Ozone data

5.1 Regression and bootstrapped variable selection

The explanatory variables of the Ozone data set with selection frequencies are listed in Table 3.

Regression and bootstrap variable selection have been carried out as described in Sect. 3.1, but we have used twice 500 bootstrap replications using each of AIC and BIC as stopping criteria for variable selection by backward elimination, and we will be interested in the extent to which these deliver different results. Furthermore, we added the models that were produced by applying backward elimination using AIC and BIC, respectively, to the full data set. Different from the Coleman data set, here these models were not found on any bootstrap sample, so there were $B = 1002$ models considered overall, with $B^* = 798$ different models. The variables SEX, FLGROSS

Table 3 Explanatory variables of Ozone data

	Name	Description	AIC sel.	BIC sel.
x_1	ALTER	Age (years)	270	117
x_2	ADHEU	Allergic rhinitis	173	44
x_3	SEX	0 male, 1 female	500*	500*
x_4	HOCHOZON	Lives high ozone village	421*	193
x_5	AMATOP	Maternal atopy	107	19
x_6	AVATOP	Paternal atopy	144	31
x_7	ADEKZ	Eczema	107	16
x_8	ARAUCH	Tobacco smoke exposure	93	24
x_9	AGEBGEW	Weight (g) at birth	183	23
x_{10}	FSNIGHT	Cough at night / morning	115	41
x_{11}	FLGROSS	Height (cm) at pfm	500*	500*
x_{12}	FMILB	Allergen sensitization	286*	131
x_{13}	FNOH24	Max. NO_2 before pfm	458*	254
x_{14}	FTIER	Animal dander sensitization	129	52
x_{15}	FPOLL	Pollen sensitization	286*	166
x_{16}	FLTOTMED	Number of medications	318*	136
x_{17}	FO3H24	Max. O_3 before pfm	281*	63
x_{18}	FSPT	Sens., any of (x_{12}, x_{14}, x_{15})	153	62
x_{19}	FTEH24	Max. temperature before pfm	278*	85
x_{20}	FSATEM	Shortness of breath	349*	195
x_{21}	FSAUGE	Itchy or watery eyes	73	6
x_{22}	FLGEW	Weight (kg) at pfm	500*	500*
x_{23}	FSPFEI	Chest wheezing or whistling	426*	288*
x_{24}	FSHLAUF	Cough following exercise	120	21

The "sel." columns indicate how often the variables were selected in the $2b = 2 * 500$ bootstrap replicates by AIC and BIC (500 replicates each)
pfm, Pulmonary function testing
Variables with "*" were selected in the original data set by AIC ($R^2 = 0.662$) or BIC ($R^2 = 0.642$)

and FLGEW were selected in all 1002 models, but otherwise model uncertainty is quite high.

On the full data set, the AIC selects 12 variables and the BIC selects 4, see Table 3, three of which were selected in all 798 models, 585 of which include x_{23}, the fourth variable selected by the BIC.

The AIC and BIC models from the original data were both only selected this one time; they did not come up exactly in any bootstrap replication. The three variables selected in all models have a strong influence on the fits and are an important reason that the fits are rather similar, see Sauerbrei et al. (2015) for related investigations. Consequently, the measure $s^* = 0.118$ shows that the variable selection instability of fits is substantially lower than for both the Coleman and the Myeloma data set.

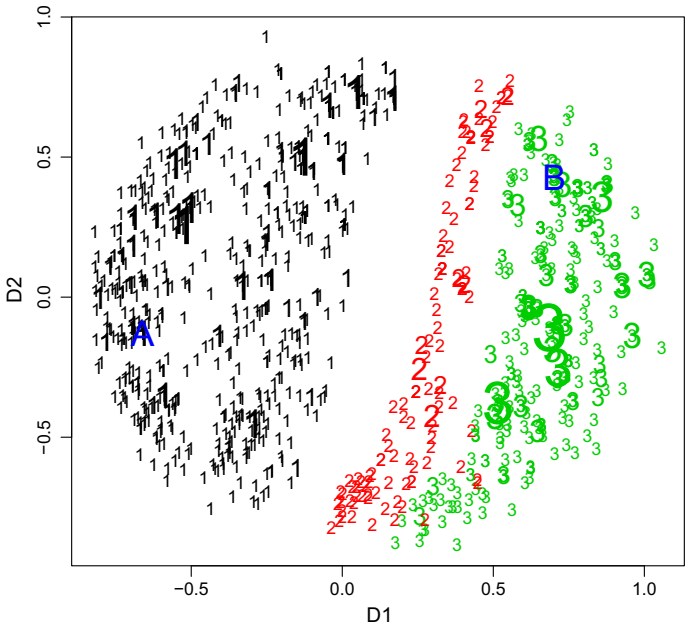

Fig. 7 Fit-based MDS for models for Ozone data. Numbers and colors indicate the 3-cluster partition from average linkage. The size of the numbers is proportional to the square root of how often the models were found. "A" and "B" denote the models found on the full data set by AIC and BIC, respectively (color figure online)

5.2 Multidimensional scaling

Figure 7 shows the 2-dimensional MDS solution for d_F. The stress is $S = 0.286$, which is not particularly good, but in terms of visible and interpretable structure higher dimensional MDS solutions do not deliver much more. There is an obvious gap separating two groups of models from each other and another not so clear gap that splits the group on the right side up into two clusters. The 3-cluster partition obtained from fit-based average linkage clustering indicated by the numbers 1, 2, 3 in the plot corresponds nicely to this. In order to investigate the meaning of this structure, we had a look at the differences between models found by the AIC and the BIC, as indicated by different colors in Fig. 8. These are strongly related to, but not completely aligned with the split between cluster 1 (mostly AIC) and the union of clusters 2 and 3 (mostly BIC). In any case it is clear from the plots that AIC and BIC select systematically very different models here, with AIC selecting models with more variables. Symbol sizes in the two plots in Fig. 8 show the rankings of the models according to the AIC (left side) and BIC (right side). These show that the AIC and the BIC disagree quite generally in this data set, with the good AIC models in the lower left and the good BIC models in the upper right of the two plots, although the models at the outer margin are rather bad according to both criteria. The good AIC models seem to occur in groups, and models that are further away from the lower left of the plot mostly yield a clearly worse AIC. The good BIC models are more scattered and good models occur in all regions of the plot that are sufficiently densely populated by BIC models.

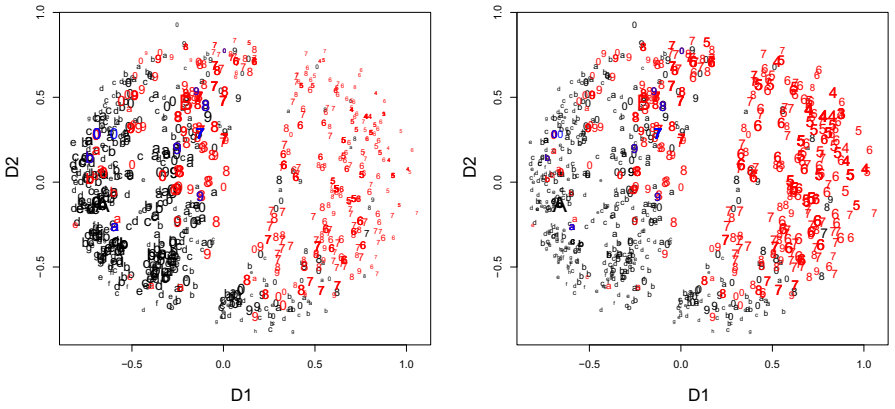

Fig. 8 Fit-based MDS for models for Ozone data. Red: found by BIC, black: found by AIC, blue: found by both (potentially on different bootstrap samples). "A" and "B" denote the models found on the full data set by AIC and BIC, respectively. Symbols show how many variables are in the models ("0","a", "b",…refer to 10, 11, 12, …variables). Left side: the size of the symbol shows the ranking of the models in terms of the AIC, i.e., the biggest symbol corresponds to the best AIC. Right side: same for the BIC (color figure online)

Furthermore, Fig. 8 explores the model sizes, i.e., the numbers of selected variables. Again there is no complete alignment although the biggest models tend to occur in cluster 1 and yield typically a higher AIC, and the smallest models tend to occur in cluster 3, connected to the BIC.

A number of decisions has to be made for producing these plots, including the range of symbol sizes and the assignment of the characteristics to different aspects of the plot (color, symbol, size). For practical exploratory analysis it is probably better to produce more plots and to focus on one or two characteristics in each plot; in Fig. 8 and later in Fig. 13 we visualized "criterion by which a model was chosen", "AIC ranking" (or BIC) and "number of variables in the model" in a single plot for reasons of space.

5.3 Clustering and heatplots

Figure 9 shows a heatplot of variables in models with the fit-based average linkage clustering. All models include the three dominating variables FLGEW, FLGROSS, and SEX. This plot characterizes cluster 1 (black, 502 out of 798 models; found on average 1.19 times) as models that all include both the variables HOCHOZON and FNOH24. The models in cluster 2 (red, 102 out of 798 models; found on average 1.17 times) include FNOH24 but not HOCHOZON, and the models in cluster 3 (green, 194 out of 798 models; found on average 1.48 times) never include FNOH24 and include HOCHOZON only very occasionally.

Heatmaps of residuals with dendrograms of models (fit-based) and observations are given in Figs. 10 and 11. Figure 10 suggests that despite the large variation in terms of the selected variables, the models actually produce quite similar fits (this is confirmed by the not shown heatplot of fits; it is in line with Sauerbrei et al. (2015), and also with the lowest value of s^* in the data sets examined in this paper). A grouping of observa-

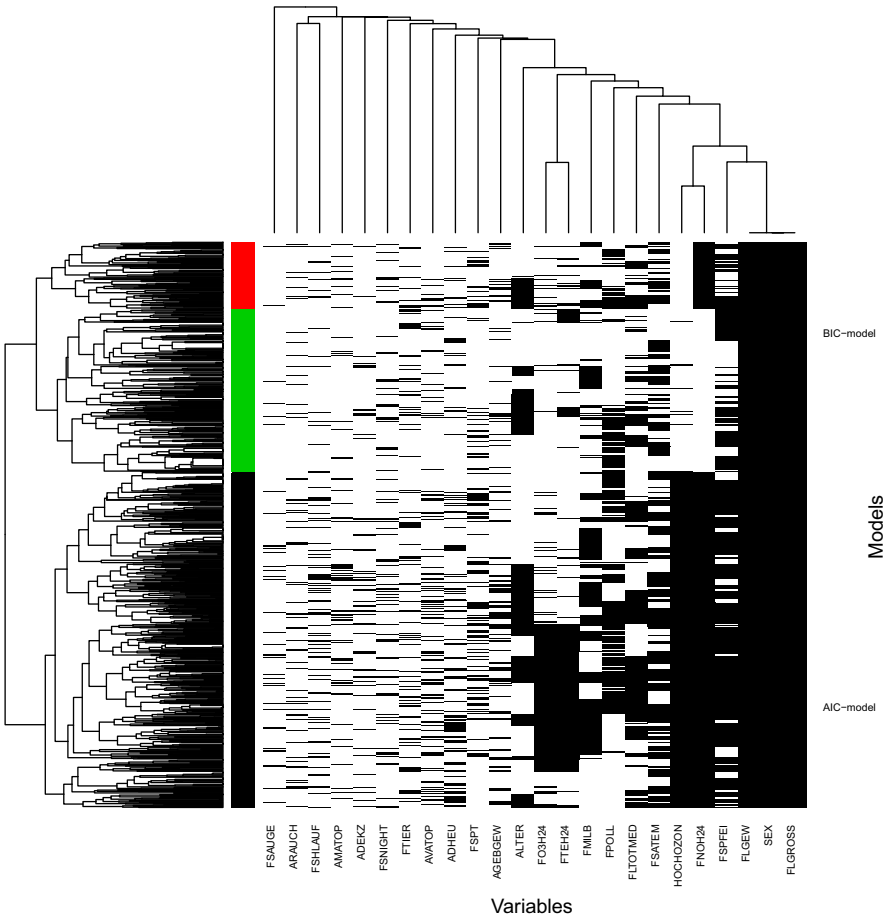

Fig. 9 Variables in models for Ozone data; model clustering dendrogram from fit-based distance. "AIC-model" and "BIC-model" denote models found on the full data set by AIC and BIC, respectively. Colors on the left side correspond to clusters from Fig. 7 (black—1, red—2, green—3) (color figure online)

tions into those that tend to produce negative residuals and those that tend to produce positive residuals with some that produce a residual around zero by almost all models seems much clearer from the plot than the clustering of the models. If interested in specific observations, one could identify and interpret the clusters of observations, but we do not do this here. Differences between the model clusters are almost invisible. A conclusion from this is that regarding the fits it matters little which model is actually chosen. The fits are quite stable; what is unstable is the selection of variables, which therefore should not be over-interpreted in any finally selected model.

Figure 11 shows column-standardized residuals, which show which observations are fitted with rather high or low values in comparison by the different models. This plot allows to see that and how the fits in clusters 1 (black), 2 (red) and 3 (green) are systematically different, with some lower variation in residuals in cluster 1 (which is connected to AIC-selection). Some cluster structure on lower levels, albeit quite weak, is also visible.

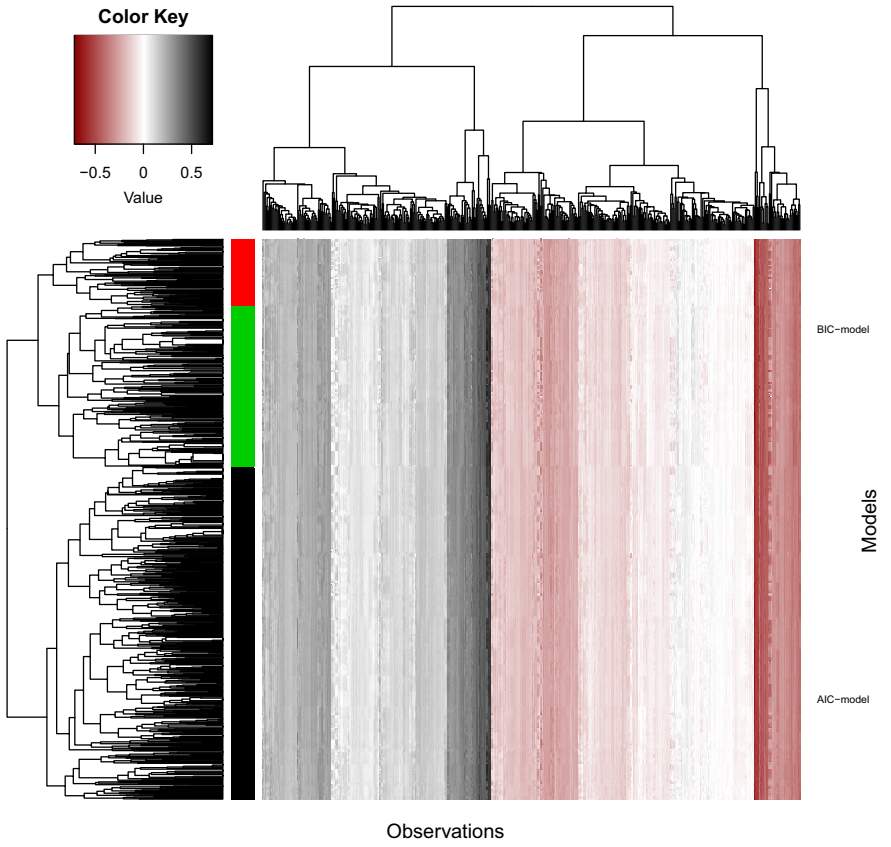

Fig. 10 Heatplot of residuals of observations versus models with fit-based hierarchical clusterings for Ozone data. "AIC-model" and "BIC-model" denote models found on the full data set by AIC and BIC, respectively. Colors on the left side correspond to clusters from Fig. 7 (black—1, red—2, green—3) (color figure online)

The most important overall message regarding this data set is that there is far more stability in fits than in the collection of selected variables. One can distinguish roughly two or three different ways of fitting the data, which are connected to whether the variables HOCHOZON and FNOH24 are in the model or not. The models selected by AIC and BIC differ strongly with AIC selecting bigger models that belong mostly to cluster 1.

6 Myeloma data

6.1 Regression and bootstrapped variable selection

Variable selection is again done using backward elimination guided by the AIC and the BIC, respectively. Because of the small sample size the difference of the penalty factors for AIC and BIC is smaller than in the Ozone data [we used the log of the number of

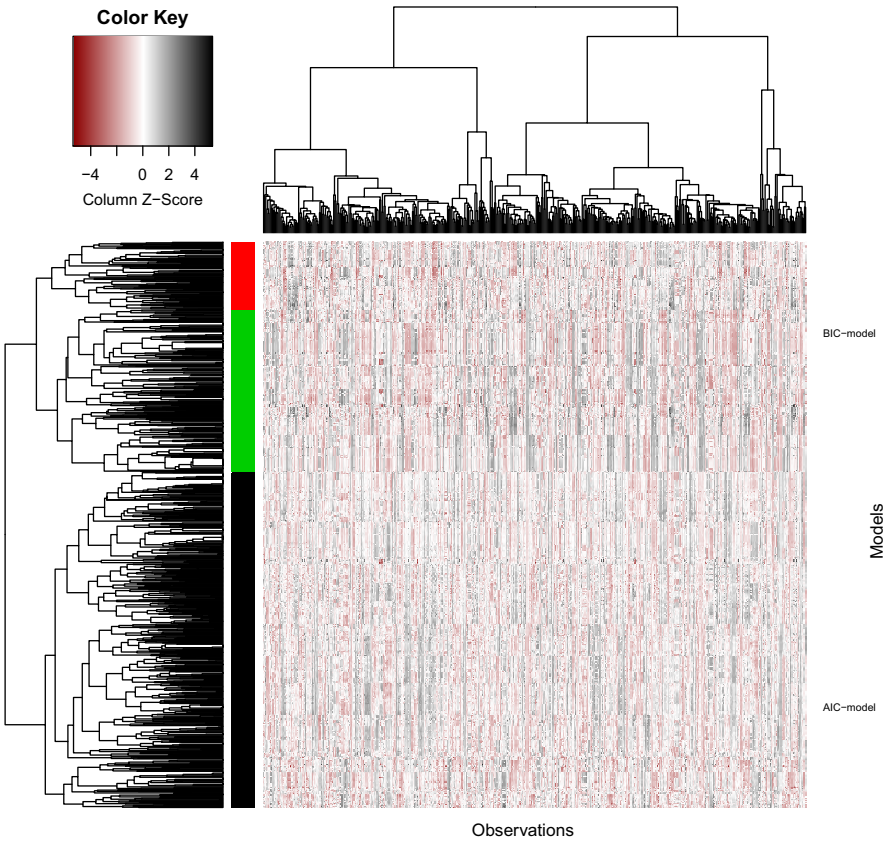

Fig. 11 Heatplot of residuals of observations versus models with fit-based hierarchical clusterings for Ozone data. Values are column-standardized, i.e., all column-wise means are zero and standard deviations are 1. "AIC-model" and "BIC-model" denote models found on the full data set by AIC and BIC, respectively. Colors on the left side correspond to clusters from Fig. 7 (black—1, red—2, green—3) (color figure online)

events here as penalty for the BIC, as recommended in Volinsky and Raftery (2000)], and consequently models selected with AIC and BIC are more similar.

As in Sect. 5, we produced $B = 1002$ models (twice 500 bootstrapped data sets and the AIC- and BIC-selected model on the full data set) of which $B^* = 780$ were pairwise different. Table 4 gives the selection numbers of the variables, and the variables selected on the full data set by AIC, BIC, respectively. Actually on the full data set the two selected models are the same, with 8 variables. However, as expected, BIC selected smaller models than AIC in many bootstrap replications (BIC selected on average 7.5 variables, AIC 9.9).

In the Cox proportional hazards model there are various ways to define fits and residuals. For the definition of the fit-based distance d_F we use as fits the expected number of deaths per month given $x_1, \ldots, x_p$; we chose these here rather than the linear predictor because the expected number of deaths is directly interpretable in practice. As residuals for the heatplots we use martingale residuals, which arise from comparing the death indicator with the expected number of deaths after the survival

Table 4 Explanatory variables of Myeloma data

	Description	AIC sel.	BIC sel.
x_1	Log BUN at diagnosis	420*	368*
x_2	Hemoglobin at diagnosis	286	228
x_3	Platelets at diagnosis	348*	278*
x_4	Infections at diagnosis	330*	245*
x_5	Age at diagnosis	229	144
x_6	Sex	334*	263*
x_7	Log WBC at diagnosis	381*	308*
x_8	Fractures at diagnosis	328*	243*
x_9	Plasma cells in bone marrow	190	108
x_{10}	Lymphocytes in peripheral blood	153	78
x_{11}	Myeloid cells in peripheral blood	185	108
x_{12}	Proteinuria at diagnosis	417*	339*
x_{13}	Bence Jone protein in urine	450*	395*
x_{14}	Total serum protein at diagnosis	351	272
x_{15}	Serum globin (gm%) at diagnosis	287	196
x_{16}	Serum calcium (mgm%) at diagnosis	247	173

The "sel." columns indicate how often the variables were selected in the $2b = 2 * 500$ bootstrap replicates by AIC and BIC (500 replicates each)
Variables with "*" were selected in the original data set by AIC or BIC (same model: Cox and Snell pseudo-$R^2 = 0.390$)

time (equal to the fit times the survival time) of the patient, see Therneau et al. (1990). The asymmetry of the distribution of martingale residuals is not an issue, because the diagnosis of the model assumptions is not our main aim.

$s^* = 0.309$ is the largest value among the data sets analysed here. The variable selection instability looks quite substantial for this data set.

6.2 Dissimilarities, multidimensional scaling, clustering

Figure 12 shows the 2-dimensional MDS solution for d_F. The stress is $S = 0.277$; again in terms of visible and interpretable structure higher dimensional MDS solutions do not deliver much more, despite improving the stress. The left side of the plot shows the average linkage clustering with 3 clusters. Clusters 1 and 2 do not seem to be very strongly separated. There is a big red "2" indicating a model from cluster 2 at the upper left side of cluster 3, which testifies that not all dissimilarity information is properly represented in a 2-dimensional MDS. Quite a bit of heterogeneity can be seen within cluster 1. Particularly, there seems to be a very homogeneous subcluster of cluster 1, containing the model that was found on the original data set by both AIC and BIC. This cluster is highlighted in cyan on the right side of Fig. 12. In the average linkage dendrogram, this cluster is only isolated at a very low level, i.e., when partitioning into very many clusters (the cluster was isolated when asking for a partition into 70 clusters of the average linkage hierarchy). This means that in fact the isolation of this subcluster is not that strong compared with dissimilarities between other subclusters,

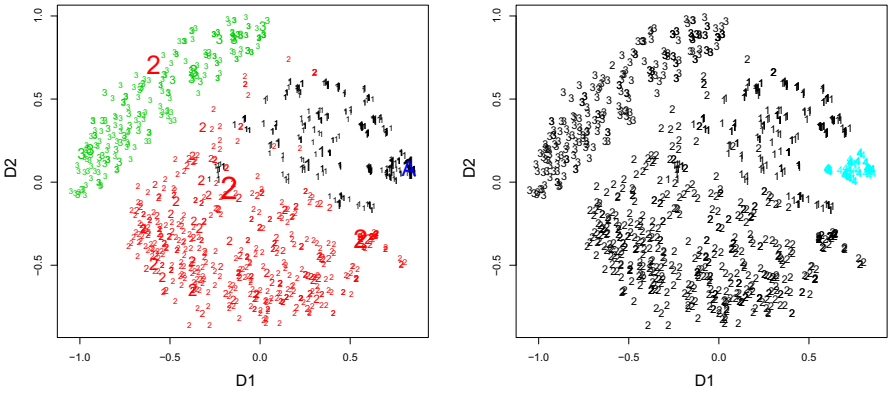

Fig. 12 Fit-based MDS for models for Myeloma data. Left side: 3-cluster partition from average linkage indicated by colors and numbers. The size of the points is proportional to how often the models were found. "A" denotes the model found on the full data set by AIC (BIC found the same one). Right side: A very homogeneous subcluster (given number 4) of cluster 1 that requires a 70-cluster average linkage-partition to be isolated (color figure online)

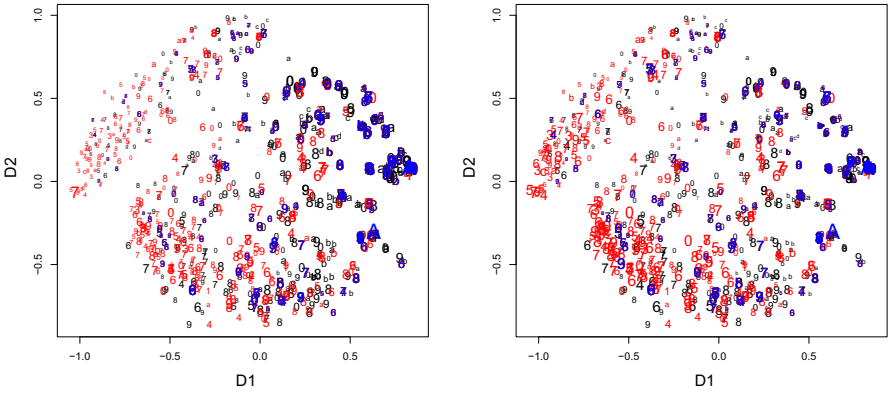

Fig. 13 Fit-based MDS for models for Myeloma data. Red: found by BIC, black: found by AIC, blue: found by both. "A" denotes the models found on the full data set by AIC (BIC found the same one). Symbols show how many variables are in the models ("a", "b",…refer to 10, 11, 12, …variables). Left side: the size of the symbol shows the ranking of the models in terms of the AIC, i.e., the biggest symbol corresponds to the best AIC. Right side: same for the BIC (color figure online)

but compared with its own homogeneity, its isolation is still strong and this makes it a potentially interesting cluster. Because cluster analysis is used here for exploratory reasons only, and we are not concerned about estimating a "true" number of clusters, in the following we will consider four clusters, namely the 3-cluster partition of average linkage but with the lower level cluster highlighted on the right side of Fig. 12 as cluster 4.

Figure 13 shows that the AIC/BIC-selected model on the full data set is surrounded by many models that were found by both AIC and BIC, with mostly many variables, in cluster 4. It also shows that the differences between the model selection by the AIC and the BIC are by far not as strong here as they were for the Ozone data in Sect. 5. The

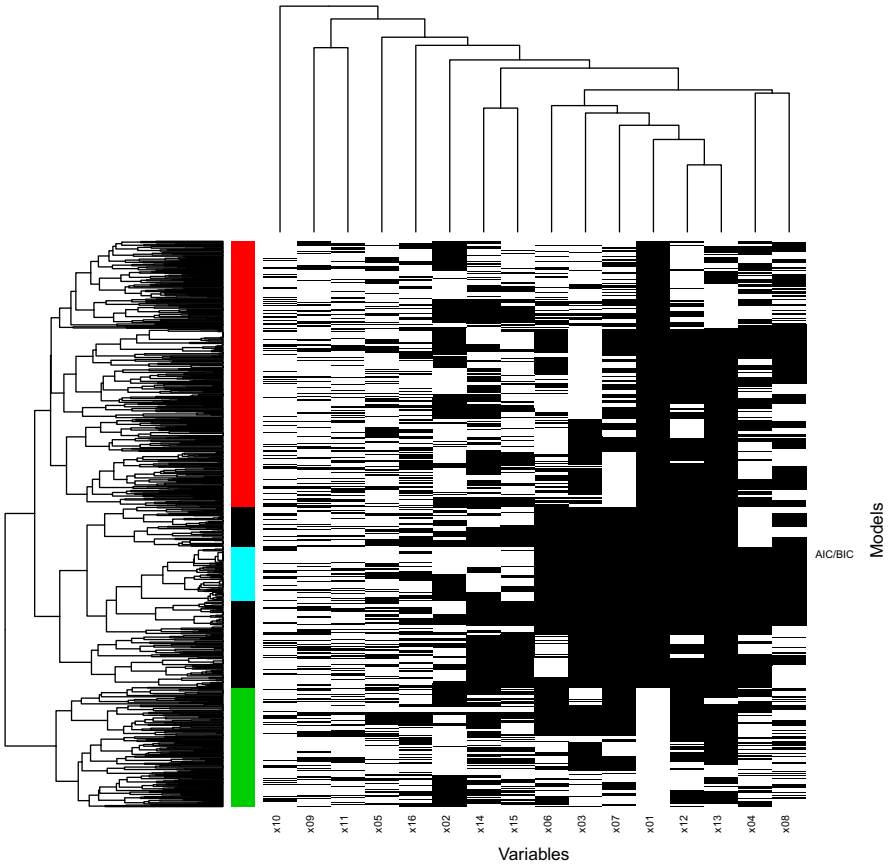

Fig. 14 Variables in models for Myeloma data; model clustering dendrogram from fit-based distance. "AIC–BIC" denotes the model found on the full data set by AIC and BIC. The color bar on the left shows the clusters from Fig. 12 (black—1, red—2, green—3, cyan—4) (color figure online)

number of models found by both criteria is much larger. This is not surprising, because using the number of events in the BIC penalty means that the BIC penalty term is not much larger than the AIC one. Bigger models tend to be on the right side of the first MDS-dimension (in practical analysis it pays off to enlarge the plots in order to see more detail particularly within cluster 4). The model clusters are not strongly related to AIC versus BIC selection. By and large, as can be generally expected, AIC-selected models are bigger than BIC-selected models, and there is a number of BIC-selected (red) models on the left side of the first MDS-dimension that are much better according to the BIC than to the AIC, but there are also many models with a quality ranking that is very similar according to AIC and BIC.

6.3 Heatplots

Figure 14 shows a heatplot of the variables in the models with the fit-based average linkage clustering. This plot characterizes cluster 3 (green, bottom, 179 out of 768

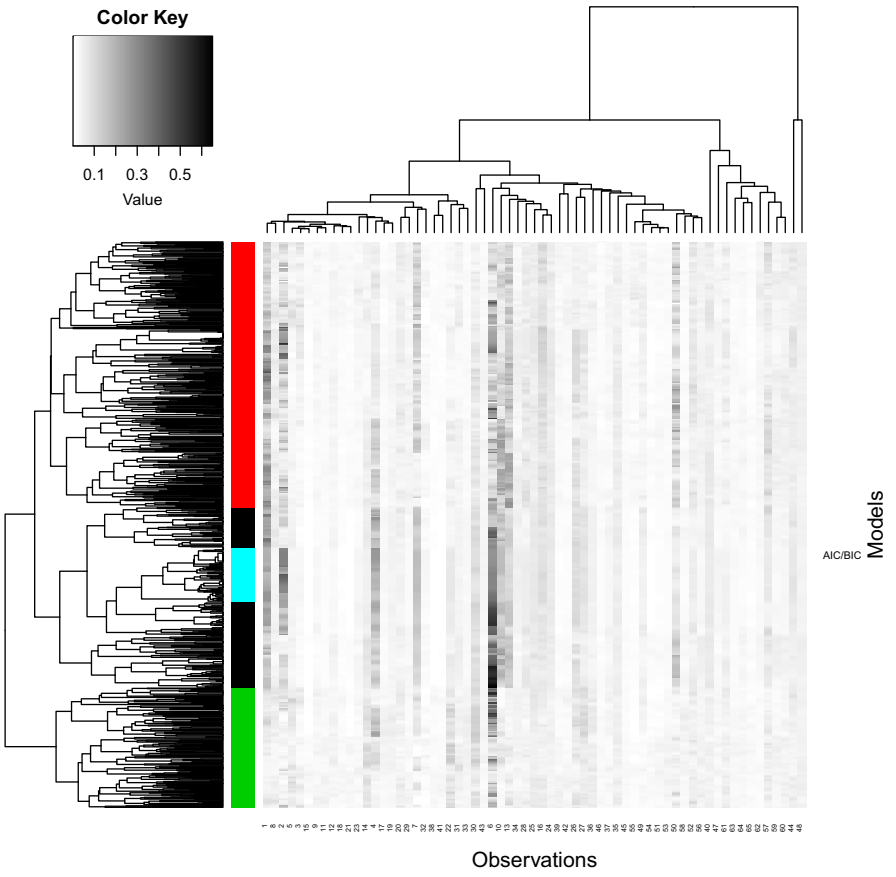

Fig. 15 Heatplot of fits of observations versus models with fit-based hierarchical clusterings for Myeloma data. The color bar on the left shows the clusters from Fig. 12 (black—1, red—2, green—3, cyan—4). "AIC–BIC" denotes the model found on the full data set by AIC and BIC (color figure online)

models, found on average 1.18 times) by the absence of x_1. Cluster 3 is merged with the other clusters at the top level of the dendrogram, meaning that these fits are the most distinct group of fits alternative to the mainstream. Cluster 4 (cyan, 50 out of 768 models, found on average 1.92 times) is characterized by the presence of all eight variables in the AIC-model selected on the full data set, plus some more. Some even bigger models of this kind are grouped in cluster 1 (black, 148 out of 768 models, found on average 1.51 times) below cluster 4. Note that cluster 4 was identified based on d_F; the variable-based d_K would not separate this group of models as a distinctive cluster, but would yield a generally less expressive clustering (not shown). The rest of the models in cluster 1 has most but not all of these variables, and additionally we often find x_{14} and x_{15}. In cluster 2 (red, 391 out of 768 models, found on average 1.20 times), only x_1 out of these is a regular appearance; x_{12}, x_{13} and x_2 appear on some lower level sub-clusters of cluster 2.

Heatmaps of fits and residuals with dendrograms of models (fit-based) and observations are given in Figs. 15 and 16. Figure 15 shows that a number of observations are

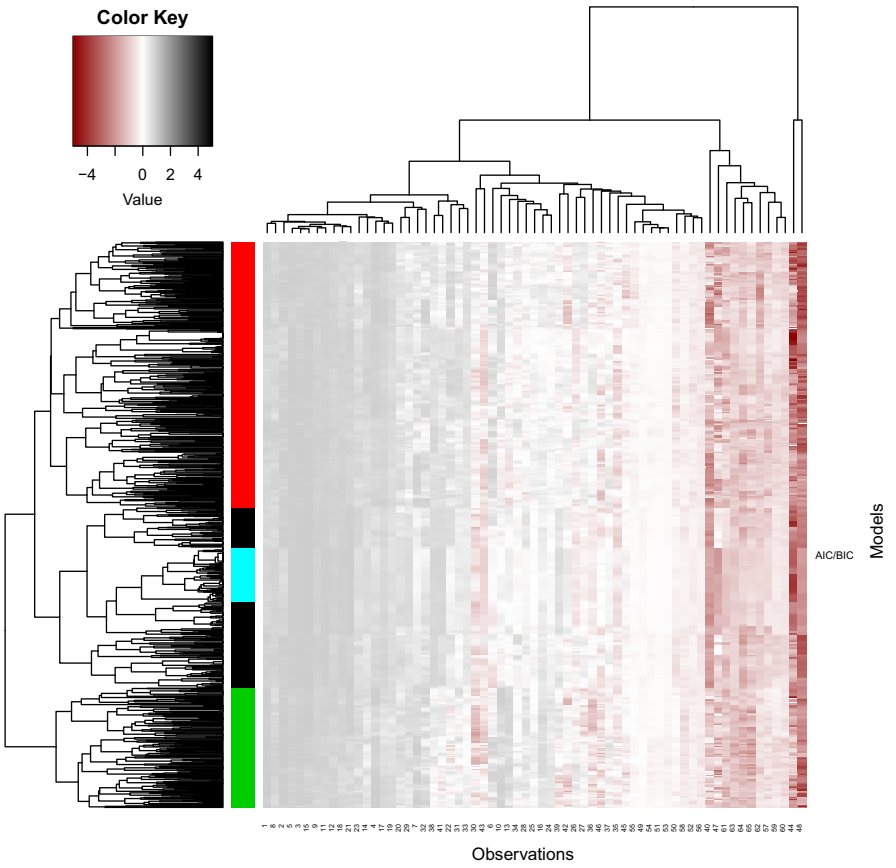

Fig. 16 Heatplot of martingale residuals of observations versus models with fit-based hierarchical clusterings. The color bar on the left shows the clusters from Fig. 12 (black—1, red—2, green—3, cyan—4). "AIC–BIC" denotes the model found on the full data set by AIC and BIC (color figure online)

fitted by the different models in quite different ways. Cluster 4 and some parts of cluster 1 seems to produce large fits (expected monthly death rates) for these observations, as opposed to clusters 2 and 3. These patients have low observation numbers, i.e., they all died quite early, and the models in clusters 1 and 4 seem more eager to fit at least some of these by use of more variables (but do not manage to assign a high death rate to all patients who died early), whereas clusters 2 and 3 produce less variation of fits. To be more precise, patients 1–6 are those who died within the first three months, and the estimated death rate per month averaged over these patients and averaged over all models within a cluster are for clusters 1–4: 0.14, 0.09, 0.08 and 0.18. Either there is overfitting in the group of clusters 1 and 4, or underfitting in clusters 2 and 3, or both. Fits in cluster 4 are very homogeneous; in the other clusters there is somewhat more variability, but the cluster structure can clearly be seen.

This is also the case for the plot of the residuals, Fig. 16. Here it can be seen that the observations on the right (which are patients with long survival times that still eventually died during the study) are hardest to fit. As in Fig. 10, there is a

clear clustering of observations, because most of them behave rather consistently over the models in terms of the residual sign. The residual structure in cluster 4 and its immediate surroundings is very stable, this is clearly and in all respects the most homogeneous cluster of models. For the other clusters, it is possible to spot groups of observations that are fitted in a stable manner in some clusters but produce a large variation of residuals in others; there is a noticeable extent of variation of residuals overall in all clusters but cluster 4.

Overall clusters 3 and 4 are the most noticeable clusters of models here. Cluster 4 has a group of very homogeneous fits similar to the AIC-selected model on the full data set, particularly assigning a large variation of fits, fitting well some of the patients who die early. The fits in cluster 3 do not vary that much; they are not strongly driven by the patients who die early, they tend to have fewer variables and in particular they exclude x_1. The remaining clusters can be seen as compromising to some extent between these. Generally the model uncertainty is rather high, and only a few observations usually make a difference when comparing fits and residuals of two models.

As in the previous examples, the fit clusters are for the Myeloma data characterized by certain patterns of included variables, rather than by model sizes, BIC- versus AIC-selection, or by the fits to specific observations.

7 Discussion

We have introduced various exploratory dissimilarity-based techniques for analysing the outcome of a bootstrap exploration of stability of model selection. These can be used to detect, for example, different groups of fits, and the sources of these differences. The difference between different model selection approaches can be explored, as well as different roles of the variables.

Using three examples we illustrate the proposed methodology and highlight some of the issues. They do not necessarily represent "typical" data sets where variable selection methods are the first choice for analysis, but they are publicly available, which allows reproducibility of our work, and the possibility to use other approaches and aim to derive further knowledge about the variability of variables selected. The Ozone data was published in an appendix of Sauerbrei et al. (2015), which also includes code for the analyses in that paper. The Myeloma data are available on https://www.imbi.uni-freiburg.de/Royston-Sauerbrei-book/index.html#datasets. They were used before in various methodological papers. The Coleman data has only 20 observations and 5 variables, which implies that there are only 32 models. We used such a small data set as initial example because details can be easier identified in plots of a size that works well in a journal. For bigger data sets researchers may want to use the flexibility of a computer screen to explore details more thoroughly. The multiple myeloma data is also small, which makes it hard to derive a suitable model with variable selection techniques (65 observations, 48 events and 16 variables). It was chosen because severe variable selection variability was expected (Sauerbrei and Antes 1992). In contrast, the structure of the ozone data ($n = 496$; 24 potential predictors) represents a typical variable selection problem. In an earlier paper (Sauerbrei et al. 2015) several issues were investigated and the current analysis can be considered as an extension with the

aim to better understand the relationships between variables and models selected with corresponding fits of the data.

Among the things that can be learnt from these analyses are (1) the discovery of a cluster structure in the selected models, lending itself to an easier interpretation of the variety of models, (2) visual analysis of how the models differ regarding residuals and fits, which may possibly lead to the discovery of substantial alternatives to the overall selected model, (3) exploration of how the differences between different variable selection criteria (here BIC and AIC) play out in the specific data set, (4) more specific issues such as the interplay between model sizes and similarities. We have also proposed a measure for the overall stability of variable selection.

In order to simplify all the information given in the various plots, a user might want to look at a low number of different models that represent the overall variability. This could be done by selecting one model from each cluster according to a quality criterion (this may be AIC or BIC but there may be other criteria relevant to the research in question).

We are well aware of problems caused by data-dependent modelling and stress that the aim of this paper is not formal inference but rather exploration. Already in the early nineties, Breiman (1992) heavily criticized the common practice to base inference on a "conditional model", ignoring uncertainty of model predictions, estimates of effects, and variance caused by model selection. Nowadays there is much literature on post-selection inference (see for example Berk et al. 2013; Efron 2014). In our work, the bootstrap is used for exploring a variety of models rather than for improving the inference based on one finally selected model.

Our proposals involve a number of decisions, such as the choices of distances between sets of variables, fit vectors, residual vectors, pairs of variables, and the MDS method. In order to investigate the sensitivity of our analyses to such choices, we did some alternative analyses, using the Euclidean distance for fit vectors, L_1-distance for residuals, exchanged the use of Jaccard and Kulczynski distance for models based on variable sets and variables based on model sets, and we tried out classical and Kruskal's nonmetric MDS for the Ozone and Myeloma data. Although a high agreement between results for the different choices would probably increase the user's confidence in the findings, it should be expected that results are to some extent affected by these decisions, as different choices often change the meaning of the analyses. We gave reasons for our original decisions and we believe that they are more appropriate than the alternative analyses, which were carried out purely for the investigation of sensitivity. For example, using the Euclidean distance for the vectors of fits will treat models that agree approximately on the fits for many observations but deviate strongly for one or two as much more different than L_1, compared to pairs of fits that deviate clearly but not extremely on all observations. We do not think that this is desirable. Indeed, the correlation between the two vectors of pairwise distances obtained from these two methods is 0.851 for the Myeloma data, the lowest values out of all correlations between vectors of distances obtained from alternative choices, all others being above 0.9. We used the cophenetic correlation (Sokal and Rohlf 1962) to compare the average linkage hierarchical clusterings obtained from the different distances. The impact of the change in distances on the resulting hierarchies is somewhat bigger, with the lowest cophenetic correlation at 0.571 (clusterings from Euclidean vs. L_1-distance between

fit vectors for the Ozone data). In this specific case this has some visible impact on the heatplots, which seem otherwise rather unaffected by reordering using dendrograms computed from alternative distances. Alternative MDS methods give images that are mostly in line with those from our preferred ratio MDS. There is some effect but this doesn't affect any of the data analytic conclusions presented above.

The proposed methods can be applied to various kinds of regression problems (as demonstrated with the survival data set in Sect. 6), various kinds of regression estimators (one could use robust ones, for example), various resampling schemes (such as robust versions of the bootstrap or subsampling), and various ways of selecting variables (here BIC and AIC, but one could also use the Lasso, for example). Note that currently the fit-based distance is defined by refitting a model that was selected on a bootstrap sample on the full data set. When applying the methodology to variable selection methods like the Lasso, this may not be suitable and the original estimates from the bootstrap sample may be used for computing fits.

The distances between models also allow for the definition of an index for model atypicality and for finding observations that are generally influential or of which the inclusion leads to atypical models. This is left to future work, as is a theoretical investigation of the variable selection instability measure.

We have used a fair amount of manual plot manipulation (e.g., by flexibly changing symbol sizes, color schemes and annotations) so that not everything presented here can be easily automated, and we encourage researchers to adapt the plots to their own needs. We provide the R-code for the analyses with some comments in order to ease the implementation which can be found in the Supplementary Material.

References

Berk R, Brown L, Buja A, Zhangv K, Zhao L (2013) Valid post-selection inference. Ann Stat 41:802–837

Borg I, Groenen P-J, Mair P (2012) Applied multidimensional scaling. Springer, New York

Breiman L (1992) The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. J Am Stat Assoc 87:738–754

Buchholz A, Holländer N, Sauerbrei W (2008) On properties of predictors derived with a two-step bootstrap model averaging approach a simulation study in the linear regression model. Comput Stat Data Anal 52:2778–2793

Cerioli A, Garca-Escudero L-A, Mayo-Iscar A, Riani M (2018) Finding the number of normal groups in model-based clustering via constrained likelihoods. J Comput Graph Stat 27:404–416

Coleman JS, Campbell EQ, Hobson CJ, McPartland J, Mood AM, Weinfeld FD, York RL (1966) Equality of educational opportunity. Office of Education, US Department of Health, Washington, DC

Cox DR (1972) Regression models and life-tables. J R Stat Soc B 34:187–220

de Leeuw J, Mair P (2009) Multidimensional scaling using majorization: SMACOF in R. J Stat Softw 31:i03

Efron B (2014) Estimation and accuracy after model selection. J Am Stat Assoc 109:991–1007

Harrell F (2001) Regression modeling strategies. Springer, New York

Hennig C (2015) Clustering strategy and method selection. In: Hennig C, Meila M, Murtagh F, Rocci R (eds) Handbook of cluster analysis. Chapman & Hall/CRC, Boca Raton, pp 703–730

Hennig C, Hausdorf B (2006) Design of dissimilarity measures: a new dissimilarity measure b etween species distribution ranges. In: Batagelj V, Bock H-H, Ferligoj A, Ziberna A (eds) Data science and classification. Springer, Berlin, pp 29–38

Ihorst G, Frischer T, Horak F, Schumacher M, Kopp K, Forster J, Mattes J, Kuehr J (2004) Long- and medium-term ozone effects on lung growth including a broad spectrum of exposure. Eur Respir J 23:292–299

Jaccard P (1901) Distribution de la florine alpine dans la bassin de dranses et dans quelques regiones voisines. Bull Soc Vaud Sci Nat 37:241–272

Krall JM, Uthoff VA, Harley JB (1975) A step-up procedure for selecting variables associated with survival. Biometrics 31:49–57

Kulczynski S (1927) Die Pflanzenassoziationen der Pieninen. Bull Int Acad Pol Sci Lett Cl Sci Math Nat B suppl. 2:57–203

Riani M, Atkinson AC (2010) Robust model selection with flexible trimming. Comput Stat Data Anal 54:3300–3312

Rouseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

Royston P, Sauerbrei W (2008) Multivariable model-building a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, New York

Sauerbrei W, Antes G (1992) Investigations on variable selection methods in regression models. In: Faulbaum F (ed) Softstat ’91—advances in statistical software 3. Fischer, Stuttgart, pp 259–266

Sauerbrei W, Buchholz A, Boulesteix A-L, Binder H (2015) On stability issues in deriving multivariable regression models. Biom J 57:531–555

Sauerbrei W, Schumacher M (1992) A bootstrap resampling procedure for model building: application to the Cox regression model. Stat Med 11:2093–2109

Shao J (1996) Bootstrap model selection. J Am Stat Assoc 91:655–665

Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. Taxon 11:33–40

Tarr G, Müller S, Welsh A (2018) mplot: an R package for graphical model stability and variable selection procedures. J Stat Softw 83:1–28

Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-based residuals for survival models. Biometrika 77:147–160

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Volinsky CT, Raftery AE (2000) Bayesian information criterion for censored survival models. Biometrics 56:256–262